**Predictive Health: ML for Disease Prognosis**

RAM PRAKASH L1 , PRAVINESH H 2

Dept of computer science and engineering

Rajalakshmi Engineering College,Tamilnadu,India

[Prakashram0820@gail.com](mailto:Prakashram0820@gail.com) 1, [Pravineshh@gmail.com](mailto:Pravineshh@gmail.com) 2

**Abstract**

Disease prediction using machine learning is the system that is used to predict the diseases from the symptoms which are given by patients or any other user.The system processes the symptoms provided by the user as input and gives the output as the probability of the disease.Naive bayes,KNN,Decision tree,Random forest classifiers is used in the prediction of the disease which is supervised machine learning algorithm. With an increase in biomedical and

healthcare data,accurate analysis of medical data benefits early disease detection and patient healthcare. By using linear regression and decision tree we are predicting diseases like Diabetes, Malaria,Jaundice,Dengue and Hypertension etc..,

**Keywords:** Disease Prediction,Machine Learning,Naive Bayes,KNN,Decision Tree,Random Forest Algorithms.

1. **INTRODUCTION**

Machine Learning is the domain that uses past data for prediction.Machine learning,a subfield of artificial intelligence on data sets to create self-learning models.The Machine Learning algorithm has two phases namely testing,training. To predict the disease from the patient body symptoms and from the history of the given patient treatment history.Basically in this phase we use different types of classifiers namely Naive Bayes Classifiers,KNN,Random forest algorithms.Using this the health care issues has been solved and the work of doctors to predict or find the patients disease has also reduced and efficient to find the disease based on the symptoms.

ML helps us to preprocess and build models to get quick and efficient ,cleaned data to deliver results faster. By using this system doctors will make good decisions related to patients' diagnosis and based on that they can give further treatment and curation for the disease ,which improves the health care service rapidly. To improve the accuracy of large data ,the existing system has unstructured data or textual data. For the prediction of disease, the existing system will be done on linear,KNN,naive bayes,algorithms[1].

With the help of developed big data analytics technology, more attention has been paid to patients disease prediction from the perspective of big data analysis,various researches have been researched and conducted by selecting the characteristics automatically from a larger number of data to improve the efficiency and accuracy of risk classification and reduction rather than the previously selected characteristics. However,those existing works are mostly considered structured data.

1. **LITERATURE SURVEY**

( Dr C K Gomathy et al,2021)[1] In disease prediction, machine learning (ML) processes people's symptoms to predict the likelihood of various diseases. Naive Bayes classifier is a supervised machine learning algorithm that calculates disease incidence to aid in early diagnosis and patient care. Combining regression and decision trees can improve the prediction of diseases such as diabetes, malaria, jaundice, dengue fever, and tuberculosis and use the physical growth of biomedical information to improve health outcomes.

( Sethi, R. S et al ,2019)[2] Today, more mobile phones are used in the world than ever before. Cell phones are everywhere and mobile technology is growing exponentially. The functionality of mobile phones allows them to provide us with services that make people's lives better. One of the services that mobile phones can provide us is digital therapy. Additionally, mobile apps are recognized to provide cost-effective healthcare solutions. Such applications provide easy and portable healthcare services for everyone. Apps like these provide users with rich experiences where they can learn more about their health and fitness. Mobile digital health apps can use a patient's symptoms to diagnose their illness. Doctors can use this information for further consultation.

(Geluvaraj et.al., 2022)[3] Research approaches disease prediction based on the prediction models and using classifiers and classification techniques. These approaches using the KNN , RFC , NB demonstrate the high accuracy of comparing one another and making them promising for healthcare applications requiring precise predictions.

(Singh, R et.al., 2019 )[5] Precise early diagnosis is crucial for improving prognosis and survival rates,spurring advancement in ai for interpreting cardiovascular data to identify risk factors and manifestation of disorders.

(S Vijayarani et al 2020)[7]Their study elaborates the data have become important for predicting clinical outcomes and extracting useful information from general medical data. Disease prediction from big medical data enables researchers to perform tasks such as classification and policies. This study focuses on the prediction of liver disease using extraction methods, specifically naive Bayes and Support Vector Machine (SVM). The performance of this method, testing time and testing time showed that the vector material was before gambling.

(S Mohan et al, 2019)[8]Cardiovascular disease is still the leading cause of death worldwide, so its prediction is important in the analysis of medical data. Machine learning (ML) has proven useful in predictive analytics using big medical data and has also been incorporated into Internet of Things (IoT) applications. Current research is only a fraction of the machine learning research into predicting heart disease. This paper presents a new method to improve the accuracy of prediction by identifying key features and using various classification methods. Our hybrid model combines random forests and methods to increase the accuracy of heart disease detection by 88.7%.

(Kumar, Y., et al,2023)[11]Artificial Intelligence (AI) improves patient care and health by using machine learning and deep learning to perform tasks such as disease diagnosis and drug discovery. The research examined AI tools used to diagnose Alzheimer's, cancer, diabetes, heart disease, tuberculosis, stroke, high blood pressure, and more, using different clinical data such as MRI and CT scans. Articles up to October 2020 are selected from archives such as Web of Science, Scopus and PubMed. The survey compares studies on metrics such as accuracy, sensitivity, specificity, AUC, precision, recall and F1 score, demonstrating the effectiveness of clinical skills and patient treatment.

(Sumana, B et.al., 2014)[17] Their research introduces a hybrid disease diagnosis model which enhances with K-means preprocess. The four stages include cleaning, selection,first and CFS, and classifier training with different models,which produce over 95% accuracy on five different medical datasets for patients' symptoms.

(Chowdhury et.al., 2021)[18] Their survey concludes that to explores that the use of advanced ML algorithms for disease prediction with 564 instance and 18 attributes using classification algorithms such as Decision Tree,KNN,Naive bayes,RAndom forest helps to achieve accuracy rate of 91% in their research and their studies improves the healthcare outcomes in resource-constrained settings like Bangladesh.

( Al-Aidaroos et.al.,2012)[19] This case study provides an overview of the workings of data mining,highlighting its accuracy and complexity. Discusses the unique characteristics of medical records,physical requirements,and various applications in analysis and this study demonstrates the effectiveness of naive bayes as a classification and demonstrates its results with different datasets.This shows that the use of NAIVE BAYES is very helpful in clinic use and performs well on wide range of clinical problems.

(Singh, A. et al,2019)[10]The important role of the heart requires accurate diagnosis and prognosis of heart disease to avoid serious consequences, including death. Because heart disease is on the rise, early diagnosis for a good prognosis is important. Machine Learning (ML) is a branch of artificial intelligence (AI) that provides support for predicting heart disease by learning from historical data. This article evaluates the accuracy of various machine learning algorithms (k-nearest neighbors, decision trees, linear regression, and support vector machines (SVM)) using the UCI repository dataset. Python, which the Anaconda (Jupyter) book uses, provides the necessary tools and libraries to increase the accuracy and precision of these predictions.

(Battineni, G et al,2020)[20] This article reviews the application of machine learning (ML) prediction models in the diagnosis of chronic diseases that cause significant global health costs and ongoing medical needs. This study examined 453 articles published in the PubMed and CINAHL libraries between 2015 and 2019 and finally selected 22 important studies. These studies propose various machine learning models such as support vector machine (SVM), logistic regression (LR), and clustering for the initial diagnosis of chronic diseases. The advantages and limitations of each method are discussed, demonstrating that there is no single best method of treatment. The findings suggest that these machine learning models are increasingly important for long-term diagnosis and are expected to play a larger role in treatment.

( Arumugam, K et al,2023)[14].This research tells that Data mining in health is a collaboration that emerges from statistical data and is very important in evaluating the benefits of health services. Heart diabetes affects people with diabetes and occurs when the pancreas cannot produce enough insulin or the body misuses insulin. Cardiovascular disease includes conditions that affect the heart or blood vessels. While many data mining algorithms exist to predict heart disease, there is no specific data for people with diabetes. In this study, the decision tree model outperformed fine-tuned, naive Bayes and support vector machine models to better predict cardiovascular disease in diabetic patients.

(Abdellatif, A.et al.,2022)[21]Cardiovascular disease is the leading cause of death worldwide, and early detection using machine learning (ML) can reduce mortality rates. However, machine learning encounters problems such as uncertainty and evaluation of data in size classes, hindering prediction accuracy. This work presents a novel approach that incorporates infinite selection (Inf-FS) to identify important features and uses enhanced weighted random forest (IWRF) for prediction and Bayesian optimization for hyperparameter tuning. Using statlog and cardiovascular data, the model achieved the highest precision and F-measurement, outperforming other composite models. Compared with previous works, the proposed Inf-FS-IWRF model improves the accuracy of parallel data by 2.4% and 4.6%.

(Deore, S., et al..2024)[22]Advances in machine learning algorithms enable early diagnosis and prediction of serious diseases. Companies are using advanced data mining tools to analyze disease pathology to help develop pharmaceutical and business strategies. This study addresses the needs of patients and drug users by focusing on diabetes and hypothyroidism in India. Using Support Vector Machine (SVM), Naive Bayes and Random Forest (RF) models, it was found that the RF model was the most effective model with an accuracy of 90% for diabetes and 98.05% accuracy for hypothyroidism. These findings help patients receive better care and reduce the risk of complications such as heart attack, stroke and vision problems.

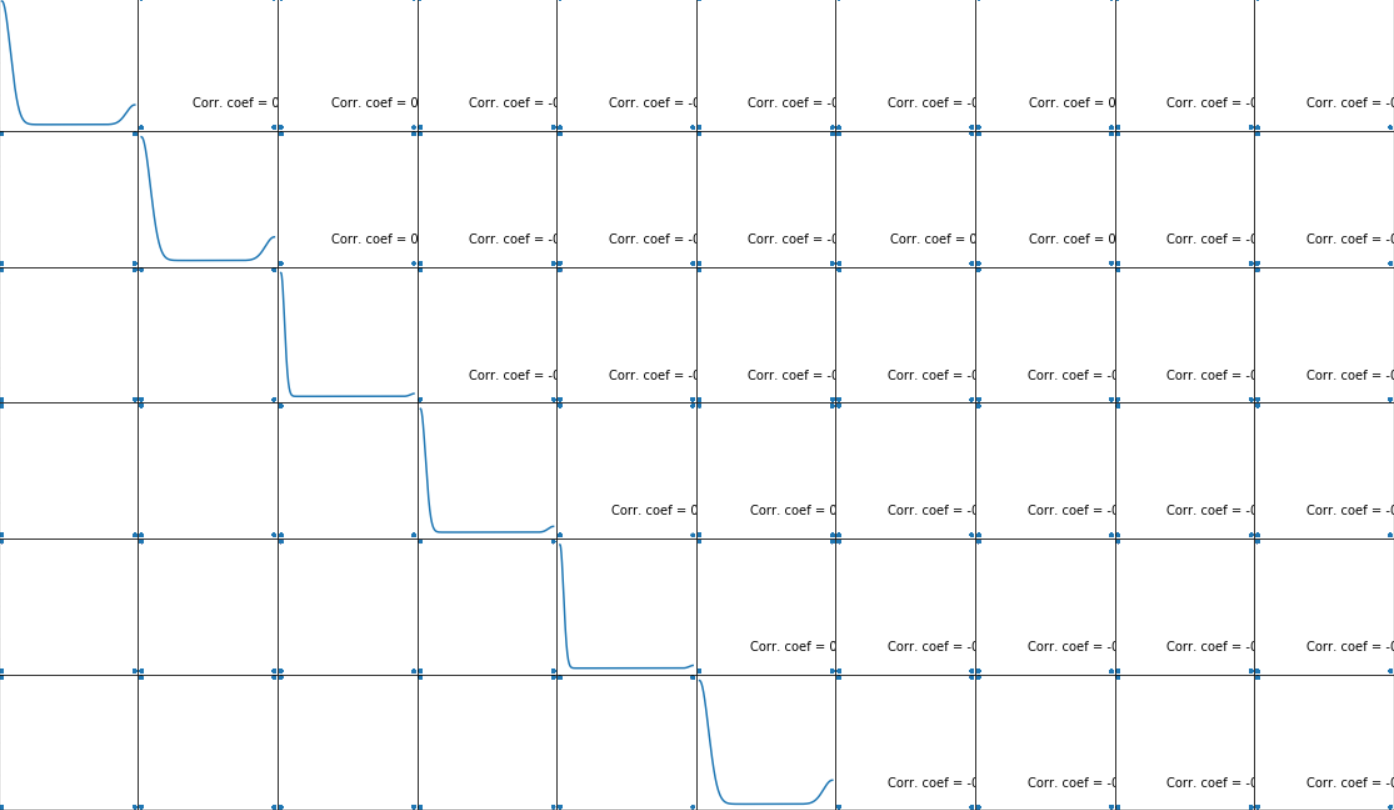
1. **PROPOSED WORK**

**DATASET ACQUISITION:** To develop an accurate learning model to predict diseases based on the patient's symptoms, it is necessary to obtain detailed information from reputable sources. This may include public health centers, research centers, hospitals and medical institutions. The primary sources of this information are the UCI Machine Learning Repository, Kaggle datasets, the National Institutes of Health (NIH) repository, and the hospital electronic health record (EHR) to ensure compliance with privacy laws. Selected data should cover a wide range of symptoms, demographics, clinical features, and comorbidities to support robust and reliable predictive models.

**METHODOLOGY:** We will follow the process to develop a powerful machine learning model to predict disease based on patient symptoms. First, we will identify and collect relevant information from reputable sources to ensure they have detailed information about symptoms, demographics, clinical features, and diagnosis. After collecting the data, we will clean and preprocess the data by handling missing values, normalizing features, and coding categorical variables. We then separate the data into training and evaluation methods to ensure the generalizability of the model. We will then use training methods to train a variety of learning models, including K-Nearest Neighbors (KNN), Naive Bayes, Decision Trees, and Random Forests. Hyperparameter tuning will improve the performance of each model. Finally, we will evaluate the model using metrics such as accuracy, precision, recall, F1 score, and AUC-ROC and select the best performing model for deployment. A user-friendly graphical user interface (GUI) will be designed to easily access the patient's symptoms and view previous diagnoses.

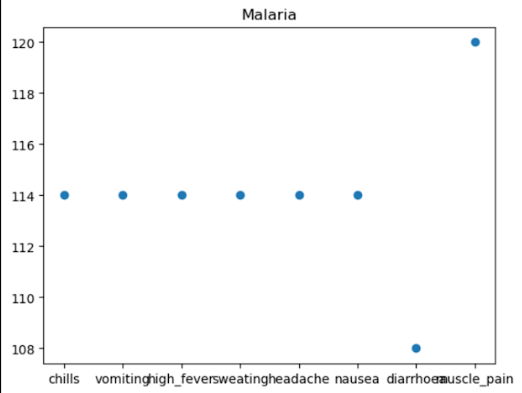
**FEATURE EXTRACTION:** Feature extraction is an important step in designing a machine learning model for disease prediction based on patient symptoms. First, we will identify the most important features from the data collected, focusing on symptoms, demographics, and clinical data that contribute to the diagnosis. Techniques such as recursive feature elimination (RFE), chi-square tests, and correlation analysis will be used to select most features. By reducing the size and removing redundant or irrelevant data, we ensure that machine learning models are efficient and effective. Careful elimination of this process will increase the accuracy and efficiency of the model and allow for greater disease prediction based on the patient's symptoms.

Correlation analysis**:** Scatter matrices and correlation coefficients help analyze the relationship between features. Similar features may be represented repeatedly and this can be resolved by reducing dimensionality such as PCA or by selecting one of the related features.



1. **ALGORITHMS**

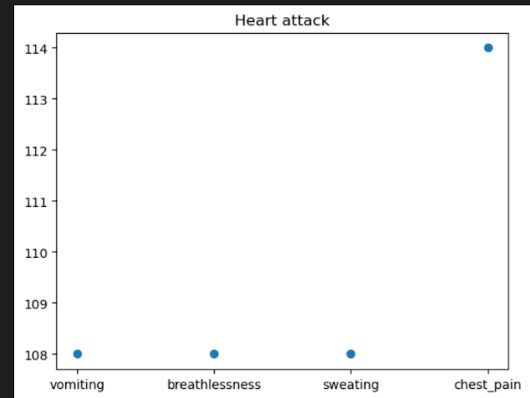
**DECISION TREE:** A decision tree is a structure that can be used to divide up a large collection of records into successfully smaller sets of records by applying a sequence of simple decision tree. With each successive division, the members of the resulting sets become more and more similar to each other. A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous (mutually exclusive) groups with respect to a particular target. The target variable is usually categorical and the decision tree is used either to: Calculate the probability that a given record belongs to each of the categories and, To classify the record by assigning it to the most likely class (or category). In this disease prediction system, the decision tree divides the symptoms as per its category and reduces the dataset difficulty**.**

****

**RANDOM FOREST:** Random Forest is a supervised learning algorithm. It is an extension of machine learning that involves packaging to improve decision trees. It combines a prediction tree and the tree is based on its own random vectors. The distribution is the same for all trees. Instead of classifying nodes based on variables, random forests use a randomly selected subset of the 12 best predictions from the nodes themselves. The worst-case learning time using random forest is O(M(dnlogn)); where M is the number of growing trees, n is the number of frequencies, and d is the data length. It is also the simplest and easiest to use algorithm. Forests consist of trees. It is said that the more trees there are, the more lush the forest will be. Random forests create decision trees based on data selected from the sample, receive predictions from each tree, and choose the best solution by voting. It also gives a good indication of what's important. Random forests have many applications, such as recommendation engines, image classification, and feature selection.

**K NEAREST NEIGHBOR:**  k-Nearest Neighbors (kNN) is a simple machine learning method. This article introduces some basic concepts of the kNN algorithm and then focuses on the use of R for kNN models. After predicting the results using the kNN algorithm, the diagnosis of the model should be checked. True mean is the most commonly used method regarding the kNN algorithm. Factors such as the k value, distance calculation, and selection of the appropriate variable value all have an impact on the performance of the model.

The nearest neighbor (KNN) algorithm is a general machine learning algorithm used for classification tasks, including disease prediction. It works by determining the distribution of unlabeled data according to the most common classes of nearest neighbors in the training program. Although classical KNN is simple, it has limitations such as equal distribution to all neighbors and distance calculation. Many different KNNs have been developed to solve these problems. These changes aim to optimize the k parameter, improve distance calculation, increase the weight of data points, and truncate the training data. Examples include adaptive KNN, locally adaptive KNN, and hybrid methods that provide multiple methods to eliminate the need to find the optimal k value. Additionally, some important variables improve accuracy through weighting, membership, or distance measurement. Each variant has a unique design and advanced performance; This highlights the importance of comparative analysis in selecting the most appropriate model for disease prediction study.[22]



**NAIVE BAYES:** Naive Bayes is an easy however amazingly powerful rule for prognostication modeling. The independence assumption that allows decomposing joint likelihood into a product of marginal likelihoods is called 'naive'. This simplified Bayesian classifier is called naive Bayes. The Naive Bayes classifier assumes the presence of a particular feature in a class is unrelated to the presence of any other feature. It is very easy to build and useful for large datasets. Naive Bayes is a supervised learning model. Bayes theorem provides some way of calculative posterior chance

P(b|a) from P(b), P(a) and P(a|b).

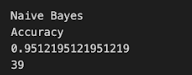
Look at the equation :

P(b v a)= P(a v b)P(b)/P(a)

Above, 11 P(b|a) is the posterior chance of class (b,target) given predictor (a, attributes) P(b) is the a priori probability of class.

P(a|c) is the chance of a predictor given the class.

P(a) is the a priori probability of a predictor.



In our system, Naïve Bayes decides which symptom is to put in the classifier and which is not. 8.3 LOGISTIC REGRESSION Logistic regression could be a supervised learning classification algorithm accustomed to predict the chance of a target variable that is Disease.

**V.EXPERIMENTAL RESULTS:** Optimal system requirements for code execution include a powerful CPU, graphics processing unit (GPU), 16 GB of RAM, and solid-state drive (SSD) storage. Python 3.6 or BASE(python 3.11.7), and it is compatible with a wide range of operating systems, including Windows, and macOS. Important libraries like NumPy, Matplotlib, scikit-learn, pandas, and tkinter must be installed before the process can begin. Installing tkinter is essential for GUI, and using Jupyter Notebook for

interactive programming is highly recommended but not required. In order to run code,

The primary sources of this information are essentials like the UCI Machine Learning Repository, Kaggle datasets, the National Institutes of Health (NIH) repository, and the hospital electronic health record (EHR) to ensure compliance with privacy laws.

**VI.** **RESULT AND ANALYSIS:** The experimental findings of the Predictive Health: ML for Disease Prognosis system, which effectively merges machine learning with Random Forest, Naive Bayes, Decision Tree, and K-Nearest Neighbors (KNN) classifiers, were acquired following a comprehensive evaluation of the model's adaptability to the dataset. The Random Forest, Naive Bayes, Decision Tree, and KNN models collectively achieved an accuracy of 95 percent in categorizing instances within the test set and navigating the complexities of prediction. This accomplishment is noteworthy.

| **METRICS** | **Classes** |
| --- | --- |
| Accuracy | 96.2 |
| Precision | 100 |

**REFERENCES:**

[1]. The prediction disease using machine learningß || Dr C K Gomathy , Mr.A.Rohith Naidu,”Volume:05” “Issue:10”||Oct 2021||IJSREM.

[2]. Sethi, R. S., Thumar, A., Jain,A. V., & Chavan, S. (2019). Disease prediction application based on symptoms. International Journal of Scientific Research in Computer Science Engineering and Information Technology, 641-646.

[3]. Geluvaraj , B., Santhosh, K., Prabhu, N., Reddy, A., Sandhya, T., & Bhaskar, S. V. (2022, January). A hybrid approach for predicting diseases using clustering and classification techniques. In *2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)* (pp. 1-6). IEEE.

[4]. Kumar, D., Balraj, K., Seth, S. *et al.* An improved machine learning-based prediction framework for early detection of events in heart failure patients using mHealth. *Health Technol.* 14, 495–512 (2024). <https://doi.org/10.1007/s12553-024-00832-z>

[5]. Singh, R., & Rajesh, E. (2019). Prediction of heart disease by clustering and classification techniques prediction of heart disease by clustering and classification techniques. *International Journal of Computer Sciences and Engineering*.

[6]. Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., ... & Mehendale, N. (2020). Disease prediction from various symptoms using machine learning. Available at SSRN 3661426

[7]. S Vijayarani , S Dhayanand Liver disease prediction using SVM and Naive bayes algorithms International Journal of Science, Engineering and Technology Research (IJSETR) , volume 4 , issue 4 Posted: 2015

[8]. S Mohan , C Thirumalai , G Srivastava Effective heart disease prediction using hybrid machine learning techniques IEEE Access , volume 7 Posted: 2019

[9]. Ramudu, K., Mohan, V. M., Jyothirmai, D., Prasad, D. V. S. S. S. V., Agrawal, R., & Boopathi, S. (2023). Machine learning and artificial intelligence in disease prediction: Applications, challenges, limitations, case studies, and future directions. In Contemporary Applications of Data Fusion for Advanced Healthcare Informatics (pp. 297-318). IGI Global.

[10]. Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In 2020 international conference on electrical and electronics engineering (ICE3) (pp. 452-457). IEEE.

[11]. Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2023). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. Journal of ambient intelligence and humanized computing, 14(7), 8459-8486.

[12]. Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making, 19(1), 1-16.

[13]. Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. Ieee Access, 5, 8869-8879.

[14]. Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T. (2023). Multiple disease prediction using Machine learning algorithms. Materials Today: Proceedings, 80, 3682-3685

[15]. Katarya, R., Meena, S.K. Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis. Health Technol. 11, 87–97 (2021).

[16]. Purvashi Mahajan,Abhishek Sharma || “Role of K-Means Algorithm in Disease Prediction” || International Journal of Advanced Trends in Computer Science and Engineering, (2016).

[17]. Sumana, B. V., & Santhanam, T. (2014, October). Prediction of diseases by cascading clustering and classification. In *2014 International Conference on Advances in Electronics Computers and Communications* (pp. 1-8). IEEE.

[18]. Chowdhury, M. N. R., Ahmed, E., Siddik, M. A. D., & Zaman, A. U. (2021, April). Heart disease prognosis using machine learning classification techniques. In *2021 6th International Conference for Convergence in Technology (I2CT)* (pp. 1-6). IEEE.[19]. Al-Aidaroos, K., Bakar, A. A., & Othman, Z. (2012). Medical data classification with Naive Bayes approach. *Information Technology Journal*, *11*(9), 1166-1174.

[20].Battineni, G., Sagaro, G. G., Chinatalapudi, N., & Amenta, F. (2020). Applications of machine learning predictive models in the chronic disease diagnosis. *Journal of personalized medicine*, *10*(2), 21

[21]Abdellatif, A., Abdellatef, H., Kanesan, J., Chow, C. O., Chuah, J. H., & Gheni, H. M. (2022). Improving the heart disease detection and patients’ survival using supervised infinite feature selection and improved weighted random forest. *IEEE Access*, *10*, 67363-67372.

[22].Deore, S., Khuniya, A., Sasi, A., Pawar, J., & Patel, M. Z. (2024). Prognostic Evaluation, Prediction and Regimen of Diseases Using SVM, NB and RF Classifiers. *Revue d'Intelligence Artificielle*, *38*(2).

[23].Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, *4*(11).

[24].Venkatesh, R., Balasubramanian, C., & Kaliappan, M. (2019). Development of big data predictive analytics model for disease prediction using machine learning technique. *Journal of medical systems*, *43*(8), 272.